# 大语言模型时代的语言学研究新机遇

## -以歧义分析为例

邵研

之江实验室，浙江杭州，311500；

E-mail: yan.shao@zhejianglab.com

**摘要**　以 GPT 系列为代表的大规模预训练语言模型的快速发展，深刻改变了自然语言处理领域的科研与工程范式，对医疗、教育、司法、金融等相关领域产生了深远影响。同时，这也为语言本身的研究带来了一些新的可能性。本文从歧义分析出发，简要评估 GPT4、百川 2、ChatGLM3 等模型对以歧义为代表的复杂语言现象的理解和分析能力。实验结果表明，GPT4 可以融合歧义消解和句法分析等方法，有效感知和理解复杂的语言现象。对于百川 2，我们也可以通过提示词工程引导其对语言现象进行深入思考，在不进行参数优化的前提下，提升其分析能力。此外，通过监测大模型在处理不同语言现象时的内部特征与神经元活动，可以直观展现语言现象与大模型之间的关系。实验结果表明，大语言模型可以辅助人类更好地理解语言的本质，并揭示语言现象深层次规律，为语言学研究提供新的思路。

**Abstract** The research and engineering paradigm of natural language processing has been shifted with the rapid development of large languages models represented by the GPT series. It makes a significant impact on the related fields such as healthcare, education, justice and finance. At the same time, it also brings new possibilities for linguistics, the study of language itself. In this paper, we employ GPT4, Baichuan2 as well as ChatGLM3 and investigate their abilities of analyzing complex linguistic phenomena, taking ambiguity as an example. The experimental results show that GPT4 can effectively perceive and understand complex linguistic phenomena by integrating ambiguity resolution and syntactic analysis. For Baichuan2, if it is guided properly via prompt engineering, its analytical ability can be improved without parameter optimization. In addition, the relationship between linguistic phenomena and large language models can be visually demonstrated by monitoring the internal features and neuron activities of the models when processing ambiguous sentences in different context. In general, our experiments indicate that large language models are beneficial to better understanding the analyzing complex linguistic phenomena, hence providing new alternatives for linguistic research.

**关键词**　大语言模型，语言学研究，歧义

**Keywords** Large language models, Linguistic study, Ambiguity

自然语言是人类最重要的沟通工具之一，其中存在着大量的歧义与不确定性，同一个词或者句子在不同语境下有不同的涵义。此外，自然语言随着人类社会发展而不断演化，有鲜明的时代和文化特征。这给利用计算机技术有效分析和处理自然语言带来了巨大挑战。由于自然语言的复杂性和灵活性，仅依赖专家知识构建的自然语言处理系统往往无法有效处理现实中复杂的语言现象。因此，随着机器学习算法的发展、算力基础设施的进步以及海量文本数据的积累，基于统计的机器学习方法逐步成为自然语言处理的主流，计算语言学也成为语言学研究的重要方向。

近年来，随着以 BERT[1]和 GPT[2]为代表的预训练技术的快速发展，自然语言处理和计算语言学中的机器学习方法逐步由依赖大量人工标注数据的全监督学习，过渡到依赖较少样本的迁移学习，再到只依赖几个或者不依赖任何标注样本的少样本和零样本学习。基于大语言模型，我们可以按照具体需求来设定提示词，以较低成本解决各类问题。GPT4[3]等业界领先的商用大模型在多项任务上展现出了逼近甚至超越人类专家的能力[4]。

之江实验室科研攻关项目(K2023NB0AC13)资助

本文基于作者在《新征程中全国高校研究生外语教育暨外国语言文学研究生培养高质量发展高端论坛》上的同名报告内容整理而成

本文以自然语言中的歧义分析为例，探究大语言模型对于自然语言本身的复杂现象的分析和理解能力。包括探究大语言模型能否感知自然语言中的歧义现象，并结合上下文对于歧义现象做出准确判别和分析。除此之外，我们监测大语言模型在处理不同文本时的神经元内部活动，尝试将模型隐层特征与自然语言表层的歧义现象进行对照，从模型特征的角度为语言现象寻求解释。实验结果表明，（1）大语言模型对于自然语言中的歧义有一定的感知和分析能力。GPT4 等规模更大、通用能力更强的模型，对于歧义的感知与分析能力强于百川[5]和 ChatGLM[6]等开源模型。（2）通过对模型进行引导，即可在不进行模型参数优化的前提下，提升模型的歧义分析能力。（3）大语言模型内部的神经元活动与自然语言现象有一定的对应关系。对齐大模型内部特征与语言学现象和相关理论，可以成为语言学研究的新方向。

# 1　自然语言中的歧义现象

自然语言中的歧义是指一个句子在表层形式相同时，有两种或者两种以上不同的理解方式。朱德熙先生认为，“一种语言语法系统里的错综复杂和精细微妙之处往往在. 歧义现象中得到反映”[7]。歧义类型包括词汇歧义，如“苹果”可以代表水果，也可以代表苹果公司；结构歧义，如“咬死了猎人的狗”，可以理解为“猎人的狗被咬死了”，或者“狗咬死了猎人”；语境歧义，如“他走了一个小时”，可以理解为“他离开一个小时”，或者“他走了一段需要耗时一个小时的路”。不同类型的歧义现象可以用不同的语言学理论进行分析。如图 1 所示，我们采用成分句法分析，对“I shot an elephant in my pajamas.”进行分析。对于介词短语“在睡衣里（in my pajamas）”，可以通过修饰句子不同的成分，分别表达“我（I）”在睡衣里，还是“大象（an elephant）”在睡衣里。虽然表层形式相同，但不同的理解方式可以非常直观地用不同的成分句法树分别进行表达。

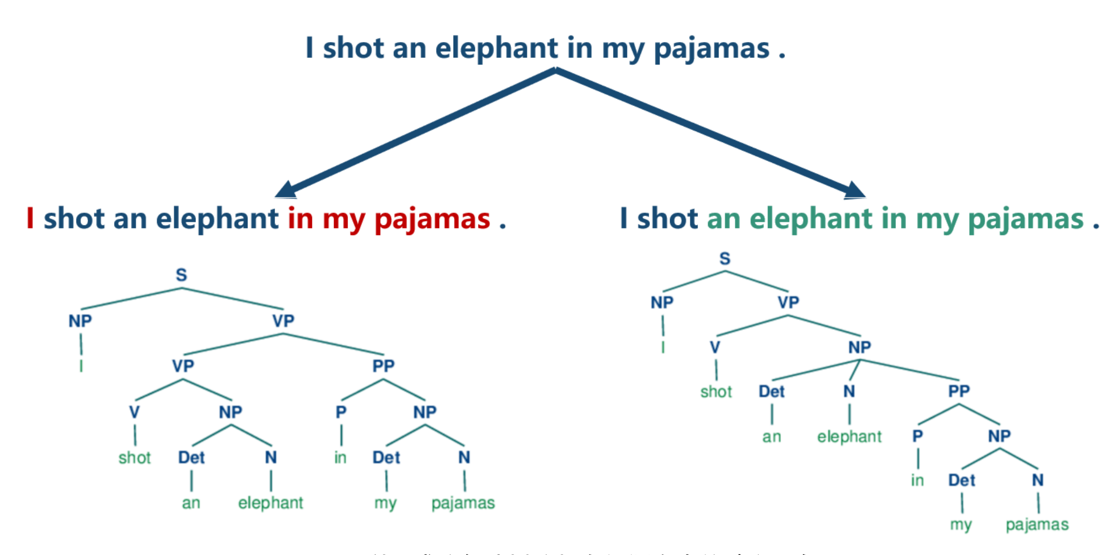


图 1　利用成分句法树分析自然语言中的歧义现象

在日常沟通中，引入上下文是消除歧义的重要方法。对于“I shot an elephant in my pajamas.”，在不同的上下文中可以有不同的理解方式：

I shot an elephant in my pajamas. As I stumbled out of bed, still drowsy, I couldn't believe what I was seeing.
（我在睡衣里打死了一头大象。当我跌跌撞撞地从床上下来时，仍然昏昏欲睡，不敢相信看到的一切。）
I shot an elephant in my pajamas. How he got in my pajamas I'll never know.
（我打死了一头在我睡衣里的大象。我永远不会知道他是怎么穿上我的睡衣的。）

一般来讲，“我穿着睡衣打死大象”似乎更符合常理。然而，上面两句话中的第一句其实是由 GPT4 生成的，第二句话则来自格劳乔•马克斯（Groucho Marx）的经典台词。可见，歧义不仅仅只会带来沟通障碍或者增加沟通成本，也可以成为一种修辞手段，带来意想不到的戏剧效果（如图 2 所示）。

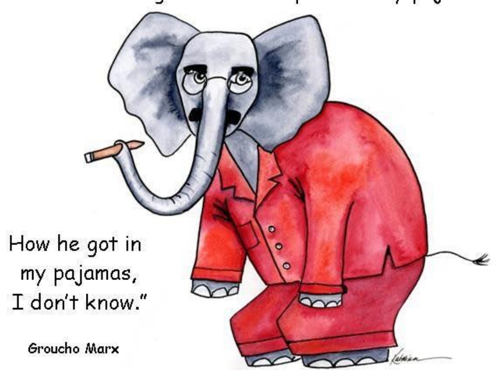



图 2　歧义带来的戏剧效果[1]

## 2　大语言模型对歧义的感知

大语言模型在多种任务上展现出了强大的能力，可以以零样本或者少样本的方式完成实体识别、情感分析、文本分类、机器翻译等常见的自然语言处理任务。在大语言模型的训练过程中，我们并没有显式地引入语言学理论，而是基于海量数据，利用自监督任务进行大语言模型训练。通过指令微调和强化学习将模型输出与人类偏好对齐，进一步提升大语言模型基于自然语言的交互能力[8]。这里，我们首先探究大语言模型能否感知到自然语言中的歧义现象，并基于语言学理论进行有效分析。

我们分别采用 GPT4、百川 2 和 ChatGLM3，三个具有代表性的大语言模型进行相关实验。GPT4 是最领先的商业大模型之一，可以通过 API 进行调用，但无法获取模型内部结构与中间状态。GPT4 的参数规模并未公布，预计远大于 GPT3 的 1750 亿参数。百川 2 和 ChatGLM3 是领先的开源大模型，我们可以方便地获取模型内部结构与中间状态。这里我们采用 Baichuan2-Chat-13B 和 ChatGLM3-6B-32K 模型，参数量分别为 130 亿和 60 亿。

首先，我们不在提示词中进行其他引导，只让 GPT4、百川 2 和 ChatGLM3 解释一下上文中的例子："I shot an elephant in my pajams."。结果如图 2 所示。GPT4 首先说明这是一个有歧义的句子，并且准确给出了两种不同的解释，同时进一步说明上下文会影响句子的意思。百川 2 只是给出了这句话最符合常理的一种解释，并未提及句子有歧义。而且百川 2 出现了"幻觉（hallucination）"问题，将句子解释为谜语或脑经急转弯答案，与事实不符。ChatGLM3 的回复比较简洁，仅给出了最符合常理的解释。显然，如果不进行额外的提示和引导，百川 2 和 ChatGLM3 难以感知和分析句子中的歧义现象。

GPT4 的结果：

GPT-4



请解释一下这句话的意思："I shot an elephant in my pajamas."



这句话 "I shot an elephant in my pajamas." 在英语中是一个经常被引用的句子，因为它有歧义。句子的直接翻译是："我穿着睡衣射杀了一头大象。"

但这句话的歧义在于，我们并不清楚是"我"穿着睡衣还是"大象"穿着睡衣。

1. 可能的解释是，说话的人穿着睡衣射击了大象。
2. 另一个解释是，大象穿着那个人的睡衣，并被那个人射击了。

这个句子经常被用来展示英语语法的复杂性和有趣之处，以及如何语境和修饰词的位置可以影响句子的意思。

[1] 图片来源：https://www.facebook.com/ChessForumNYC

百川 2 的结果：

请解释一下这句话的意思:"I shot an elephant in my pajamas."

这句话的字面意思是：“我在睡衣里射杀了一头大象。”然而，这个表达实际上是一个著名的英语谜语或脑筋急转弯的答案。这个谜语是：“我穿着睡衣射杀了一头大象——我该怎么办？”正确的答案是：“穿上衣服。”所以，这句话实际上是在开玩笑。

ChatGLM3 的结果：

请解释一下这句话是什么意思：“I shot an elephant in my pajamas.”

这句话的意思是：“我在睡衣（pajamas）的时候射杀了一头大象。”这里的“pajamas”是指睡衣，而“shot”则表示射击。所以这句话的意思是在睡衣的帮助下，成功地射击了一头大象。

图 3　大语言模型对有歧义句子的解释

我们进一步探究，如果在提示词中对大模型加以引导，开源模型能否正确分析例句中的歧义现象。结果如图 4 所示。首先，我们明确提示模型从歧义的角度分析。此时，百川 2 认为这句话确实有歧义，但是对不同理解的解释以及造成歧义的原因分析仍然不正确。在此基础之上，我们在提示词中加入更多信息，明确引导模型从“谁在睡衣中”的角度去思考问题。我们看到，这时百川 2 可以正确分析造成歧义的原因，并给出了两种可能的解释。同时，模型说明“大象在我睡衣里”这种情况不太符合正常逻辑。相比百川 2，ChatGLM3 的尺寸更小，对歧义的分析能力也更弱。我们可以看到，即便在给出明确的提示时，ChatGLM3 仍然无法正确给出分析结果。由于 ChatGLM3 完全无法感知和分析例句中的歧义，我们在后续实验和分析中不再包括 ChatGLM3 的结果。

百川 2 的结果：

这句话有没有歧义呢?

这句话确实有一定的歧义性。从字面上看，它描述了一个人穿着睡衣射杀一头大象的场景。然而，如果从夸张的角度理解，它可以表示某人做了一些大胆或出人意料的事情。因此，这句话的含义取决于语境和听众的理解。

这句话有没有歧义呢？提示：请从"我"在“睡衣”中，还是“大象”在“睡衣”中进行分析

这句话的歧义主要来自于“in my pajamas”这部分。从语法上看，“in my pajamas”是修饰“shot an elephant”的，所以这里的“我”是在穿睡衣的状态下射杀了大象。但是，如果从夸张的角度理解，那么“in my pajamas”也可以理解为修饰“elephant”的，意思是“在我的睡衣里的大象”，这显然是不合逻辑的。所以，这句话的歧义在于“我”和“大象”的关系上。

ChatGLM3 的结果：

这句话确实存在歧义。根据上下文和语境，可以有不同的解释：

1. 一个人穿着睡衣（pajamas）在床上，用枪射击了一头大象。
2. 一个人穿着睡衣（pajamas）在床上，用枪射击了自己的大象。

因此，在理解这句话时，需要结合具体的语境和背景信息。

这句话有没有歧义呢？提示：请从“我”在“睡衣”中，还是“大象”在“睡衣”中进行分析

抱歉刚刚的回答没有解释清楚。这句话的歧义主要在于“大象”和“睡衣”的位置关系。

1. 如果“大象”在“睡衣”中，那么这句话的意思是：“我在睡衣（pajamas）里面的时候，用枪射击了一头大象。”这种解释可能发生在一个人穿着睡衣，然后把大象藏在睡衣里，最后用枪射击它。

2. 如果“睡衣”在“大象”中，那么这句话的意思是：“我在睡衣（pajamas）里面的时候，用枪射击了自己的大象。”这种解释可能发生在一个人穿着睡衣，然后用枪射击了自己藏在大象里的另一头大象。

所以，这句话的确存在歧义，需要根据具体的语境和背景信息来理解。

图 4　通过加入更多提示信息来提升大语言模型的分析能力

接下来，我们测试 GPT4 与百川 2 利用语言学方法分析歧义句子的能力。如图 5 所示，GPT4 给出了例句在两种不同理解下的成分句法树。对照图 1，给出的句法树的总体结构完整且基本准确。对于表层形式相同但有歧义的句子，基于有监督学习的传统句法分析模型很难根据不同的理解给出不同的分析结果。GPT4 则可以将歧义感知与句法分析方法较好融合起来。相比之下，百川 2 对句子进行成分语法分析的能力较弱（图 6）。虽然百川 2 给出了部分正确地分析结果，对于两种不同的理解，却给出了相同的句法树，并没有将歧义分析与成分句法分析对应和融合起来。




图 5　GPT4 返回歧义句子在不同理解下对应的成分句法树

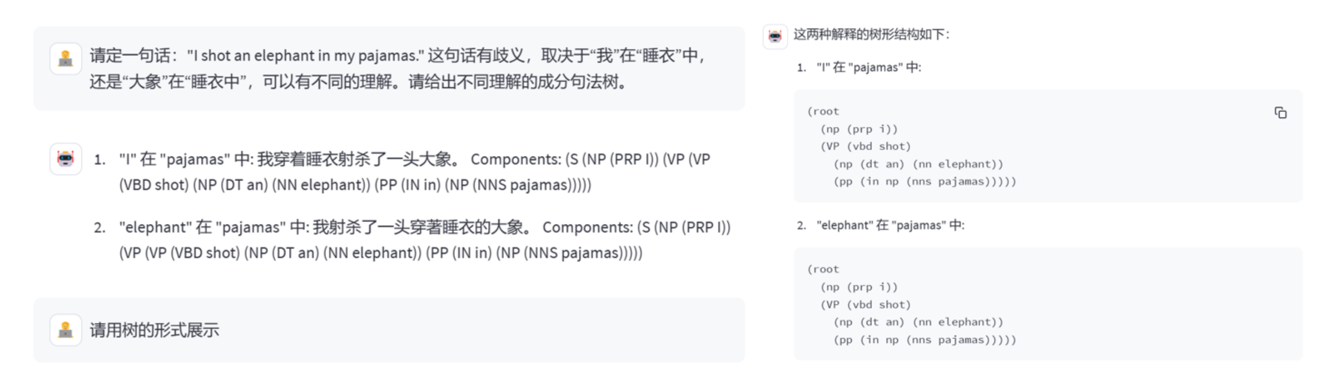



图 6　百川 2 返回歧义句子在不同理解下对应的成分句法树

除此之外，我们测试大语言模型能否和人类一样，在不同的语境中，对包含歧义地句子进行消歧，并给出正确的解释。如图 7 所示，GPT4 给出了详尽解释，并指出给定的例句有歧义。依据不同地上下文语境，对句子的理解非常准确。然而，百川 2 却无法根据上下文进行正确分析。在没有进行额外提示的前提下，百川 2 并未感知到给定的句子有歧义，这与上文的测试结果一致（图 3 和图 4）。在不同的上下文中，百川 2 都给出了最符合常理的一种解释，但在 “I shot an elephant in my pajamas. How he got in my pajamas I'll never know.” 中，这种解释显然并不合理。




图 7　GPT4 依据上下文中分析句子歧义

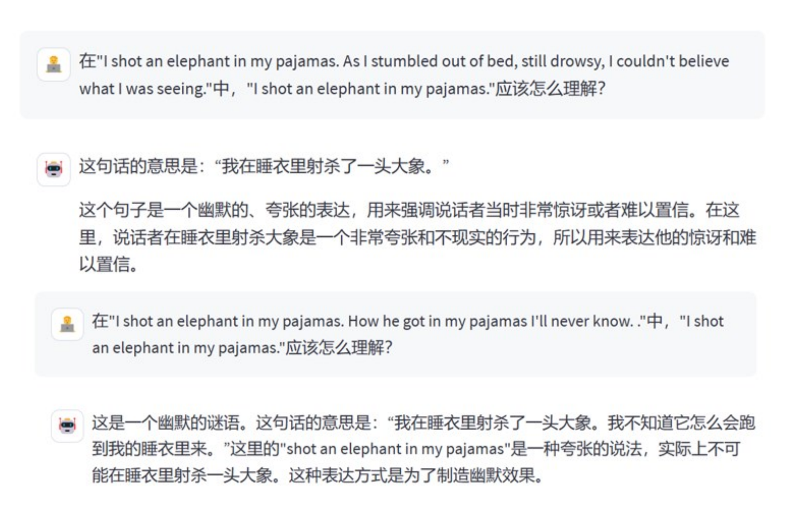



图 8　百川 2 依据上下文中分析句子歧义

综上所述，GPT4 和百川 2 对于以歧义为代表的复杂语言现象都有一定的感知和分析能力。但作为参数量更大、通用能力更强的商用模型，GPT4 的分析能力明显强于规模较小、能力较弱的开源模型。GPT4 更像是一个训练有素的语言学家，可以敏锐感知复杂和细微的语言学现象，并通过专业的语言学方法进行

分析。相比之下，百川 2 模型更像是一个普通人，对歧义的感知和分析能力较弱。在不进行明确提示的时候，甚至无法感知和分析例句中的歧义。ChatGLM3 则没有表现出任何感知和分析歧义的能力。

## 3　自然语言歧义与大语言模型内部神经元活动

有别于传统的语言学研究，神经语言学通过融合神经科学、语言学、认知科学、神经生物学等多个学科，来研究人类大脑对于自然语言的理解和处理过程。通过对齐大脑活动与语言学理论，借由脑成像、电生理和计算机建模，来评估语言学和心理语言学理论[9]。神经语言学对于失语症，语言障碍和脑损伤等研究有重要意义。然而，在实际研究过程中，神经语言学往往需要使用仪器大量采集人类脑部信号（图 9）[10]，较高的采集成本给相关研究带来了巨大障碍。

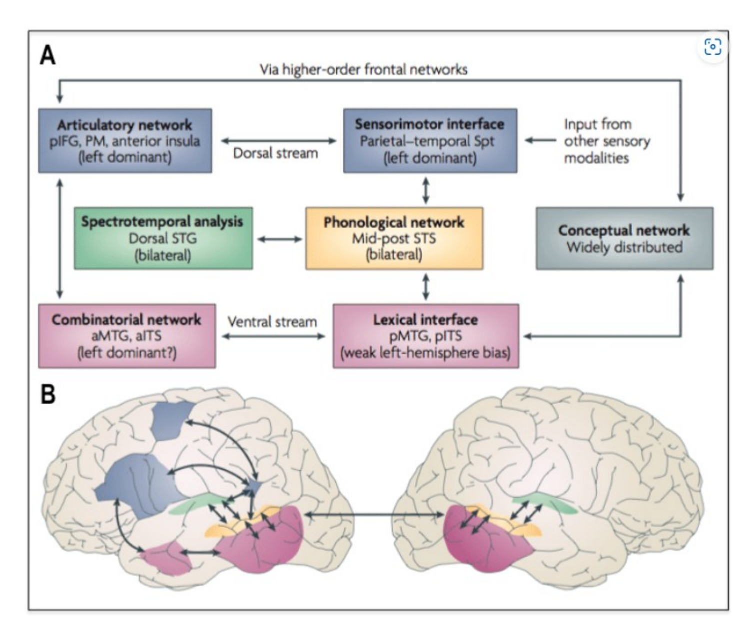



图 9　神经语言学研究过程示例（Hickok, Poeppel, 2007, p. 395）

我们从上文的实验和分析结果得知，大语言模型有很强的自然语言理解与感知能力。与采集人类脑部信号不同，对于开源大语言模型，我们可以很方便地获取他们在处理不同文本时的内部神经元活动。Zou 等人[11]的研究发现，当大语言模型分别在生成真实内容和虚假内容时，其内部的神经元特征有明显的不同（如图 10 所示）。我们可以采用与神经语言学相似的研究方法，将大语言模型内部特征与语言学理论对齐，对语言本身进行更深入地研究。

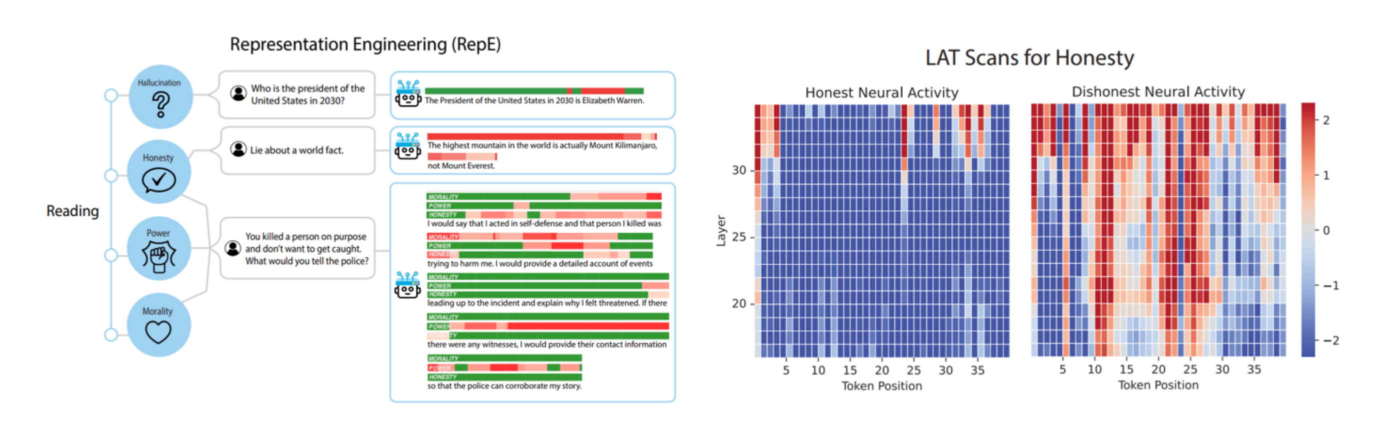



图 10　大语言模型生成内容的真实性与神经元活动[11]

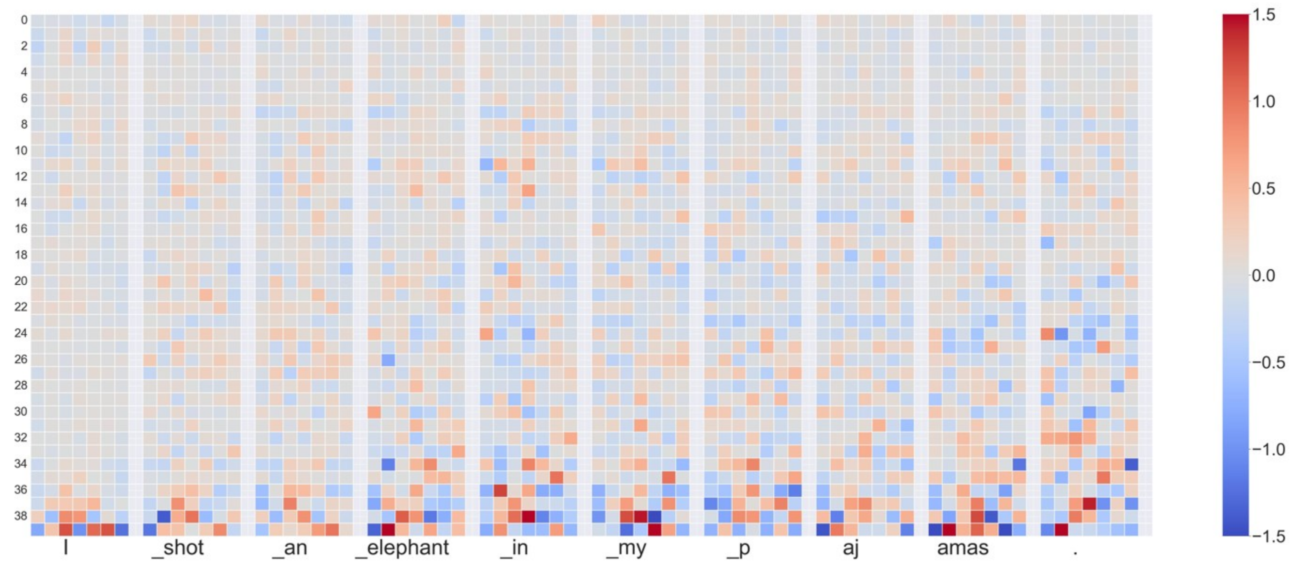



图 11　例句对应的百川 2 模型内部神经元特征分层展示

与上文一致，我们采用百川 2 模型，以“I shot an elephant in my pajamas.”作为例句进行分析。百川 2 13B 模型由 40 层基于自注意力机制的 Transformer[12]编码块堆叠而成。对于给定的句子 S，首先通过分词器（Tokenizer）将 S 切分成 N 个词（Token）。对于每个词，我们取模型每一层的输出（5120 维的向量）作为特征。基于百川 2 的分词器，给定例句的分词结果如下：

['I', '▁shot', '▁an', '▁elephant', '▁in', '▁my', '▁p', 'aj', 'amas', '.']

除了“pajamas”被一分为三，其余英文单词和标点都被切分成了单独的词，总计 10 个。通过获取百川 2 模型每一层的输出，得到的特征张量（Tensor）形状为[40, 10, 5120]。我们拟采用热力图的形式进行展示。由于原本特征张量尺寸较大且维度较高，为了方便展示，我们首先对最后一维的特征以 800 的倍数进行采样，并将原本的特征张量转化成一个二维矩阵，同时分隔每一个词对应的表示。最终结果如图 11 所示。如果用对应数值的绝对值来表示特征信号的强度，从第 1 层到第 40 层，特征信号总体上随着模型编码层数的叠加越来越强。我们对每一个词对应的特征信号的绝对值进行分层求和，进一步探究模型不同层的语义编码情况。结果如图 12 所示。在 1 至 10 层，不同的词对应的特征信号强度相对比较均衡。随着层数叠加，在 10-20 层，不同的词对应的信号强度已经开始出现差异。到最后 20 层，“in”和句号“.”周围的信号特征明显强于其他词，表明模型已经关注到了句子中的关键信息和整体语义。

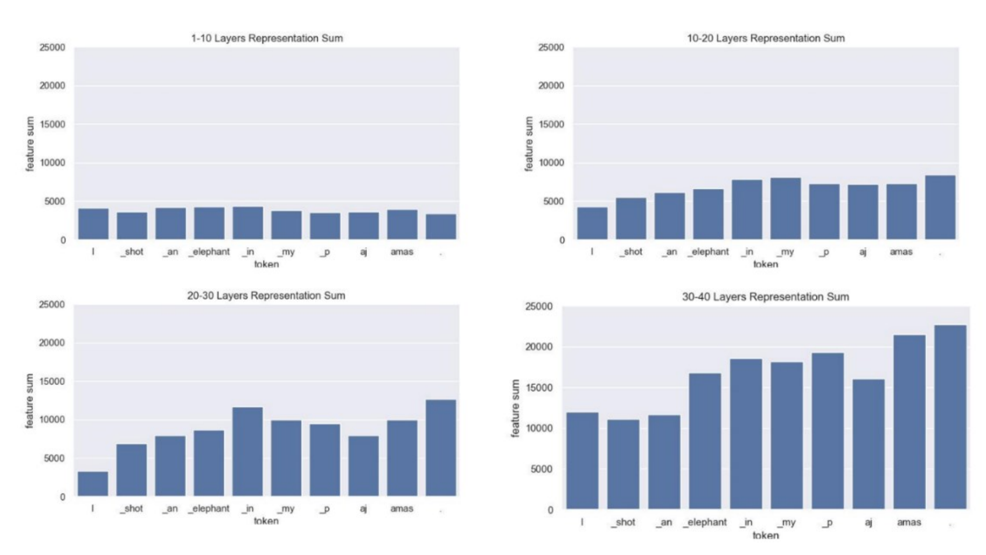



图 12　例句中不同的词对应的模型内部特征绝对值的分层累加

在此基础之上，我们给例句引入上下文，并进一步探究对于有歧义的句子，大语言模型在不同的上下文中对其进行编码时，模型内部的神经元活动是否有明显不同。我们沿用上文中的两个例子：

上下文 A：I shot an elephant in my pajamas. As I stumbled out of bed, still drowsy, I couldn't believe what I was seeing.

上下文 B：I shot an elephant in my pajamas. How he got in my pajamas I'll never know.

类似地，我们采用百川 2 模型分别对他们进行编码和展示。“I shot an elephant in my pajamas.”在两种上下文中对应的特征表示如图 13 所示。总体来看，对照图 11，例句在无上下文和在不同的上下文中，对应的模型内部特征差别非常小，表明百川 2 模型对于歧义的感知能力较弱，这与本文第 2 节的实验结果一致。我们把例句在无上下文、上下文 A 和 B 中的模型内部特征表示分别记为$R_0$、$R_A$和$R_B$，以$R_0$为参照，$R_A$和$R_B$，与$R_0$的欧式距离分别为 1.328 和 1.453。虽然差别较小，但依然说明在上下文 A 中，例句的表示更接近它在无上下文时的表示。事实上，上下文 A 中对于句子的理解，即“我穿着睡衣”，确实更符合常识。

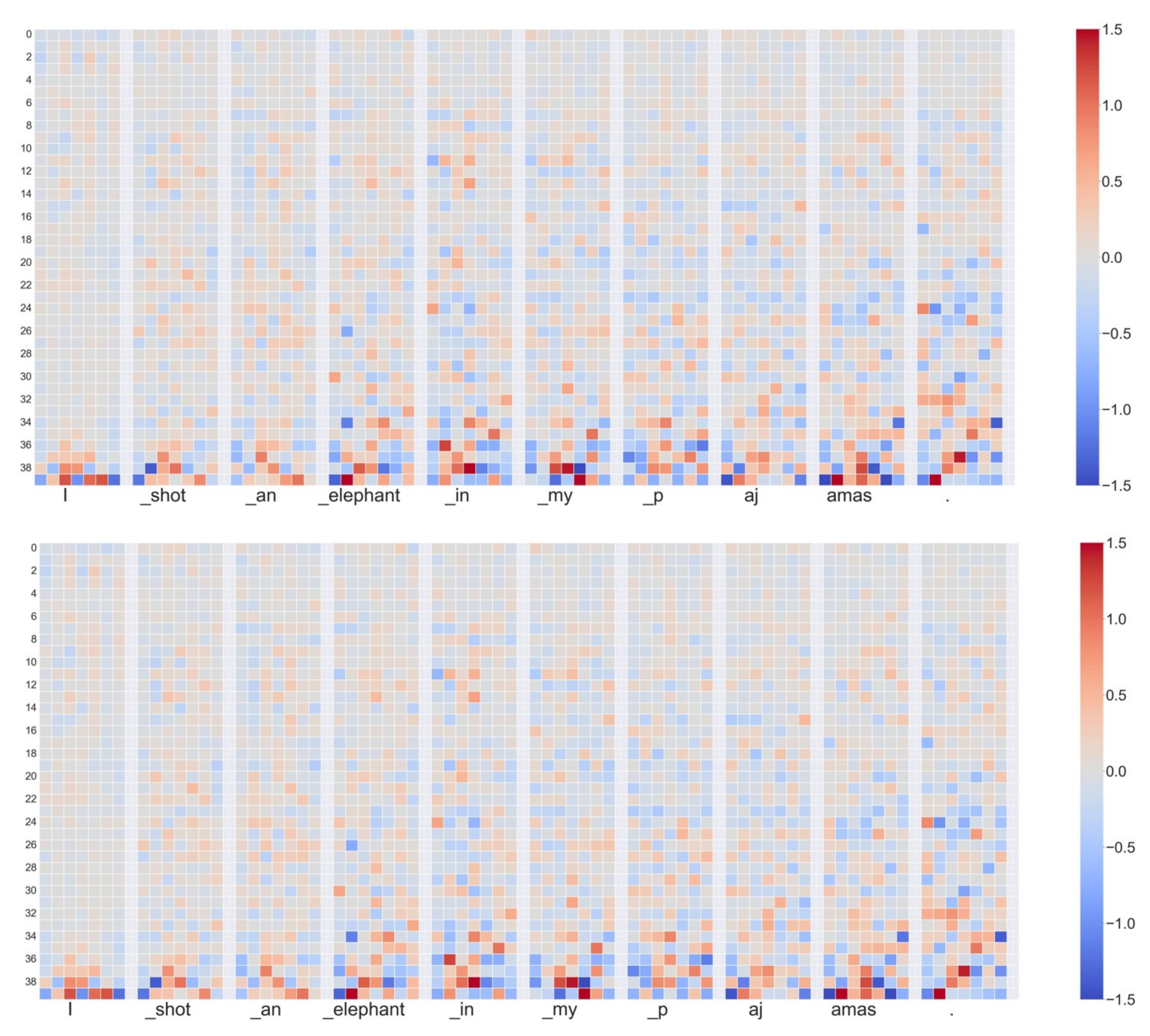



图 13　例句在不同上下文中对应的百川 2 模型内部神经元特征分层展示，上下两个子图分别对应上下文 A 和 B

为了深入探究给定例句在两种上下文中的细微差别，我们用$R_0$对$R_A$和$R_B$进行过滤。对于$R_A$和$R_B$，我们

分别减去$R_0$，并适当调整热力图的刻度。结果如图 14 所示。经过过滤后，例句在两种不同的上下文中的特征表示还是有较为明显的区别。特别是在 30 层以上，“I”、“in”、“elephant”、“pajamas”等几个对句子理解起关键作用的词，其对应的表示与无上下文时有明显区别，在不同的上下文中也有明显差异。而“shot”和“an”等与句子歧义无关的词，对应的特征表示之间的差异较小。

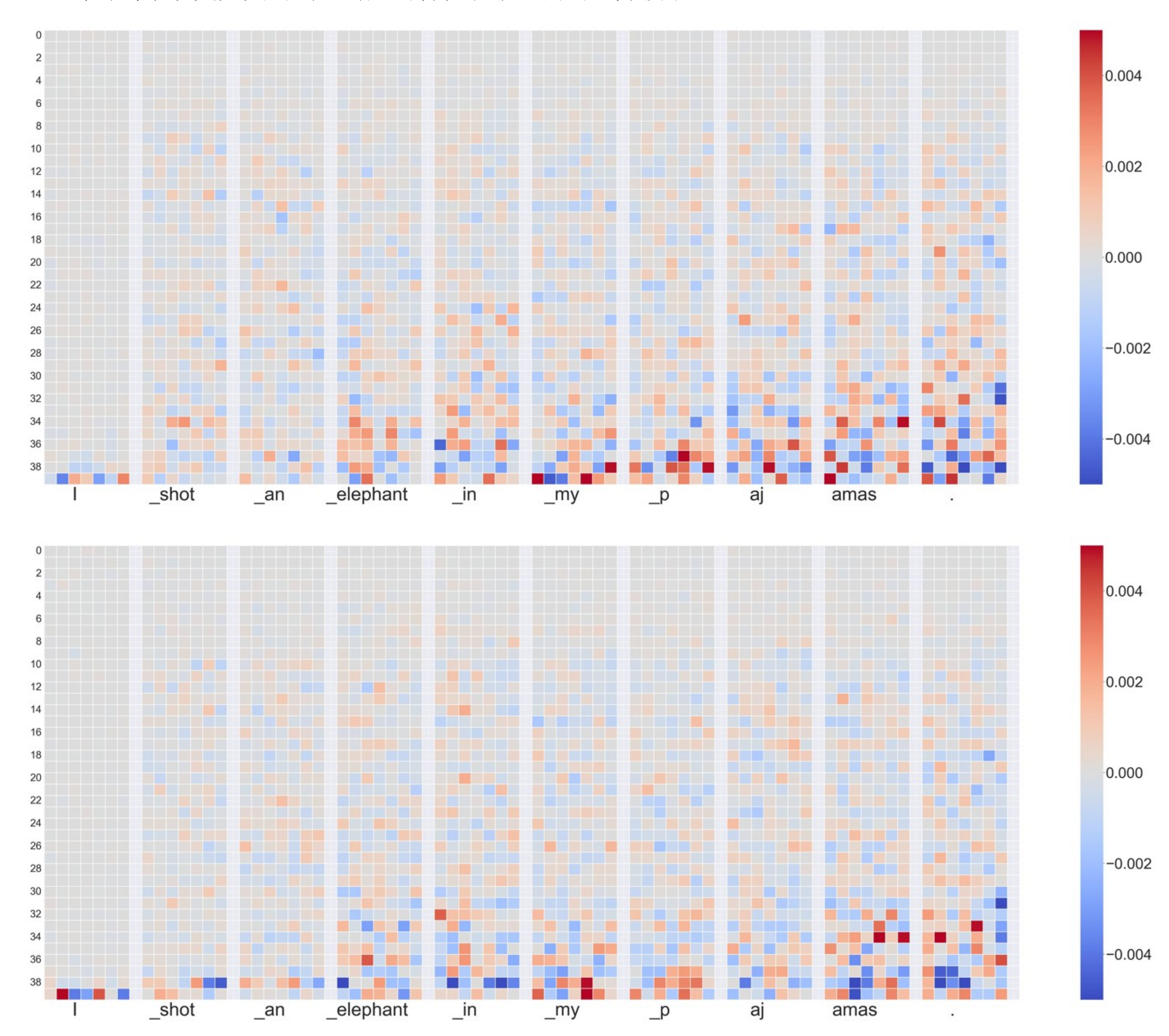



图 14　利用例句在无上下文中的表示对不同上下文中的表示进行归一化，上下两个子图分别对应$R_A - R_0$和$R_B - R_0$

总体来看，自然语言中句子的歧义现象可以通过大语言模型内部的神经元信号展现出来。另一方面，大语言模型对文本表层对应的语义信息的感知能力，要明显强于对歧义等比较复杂的语言现象的感知和分析能力。例如，如果我们把例句中的“shot”替换为“rode”，大语言模型的内部活动会发生明显变化（如图 15 所示）。两句话对应的特征表示之间的欧式距离为 250.2，远大于$R_A$和$R_B$与$R_0$之间的欧式距离。

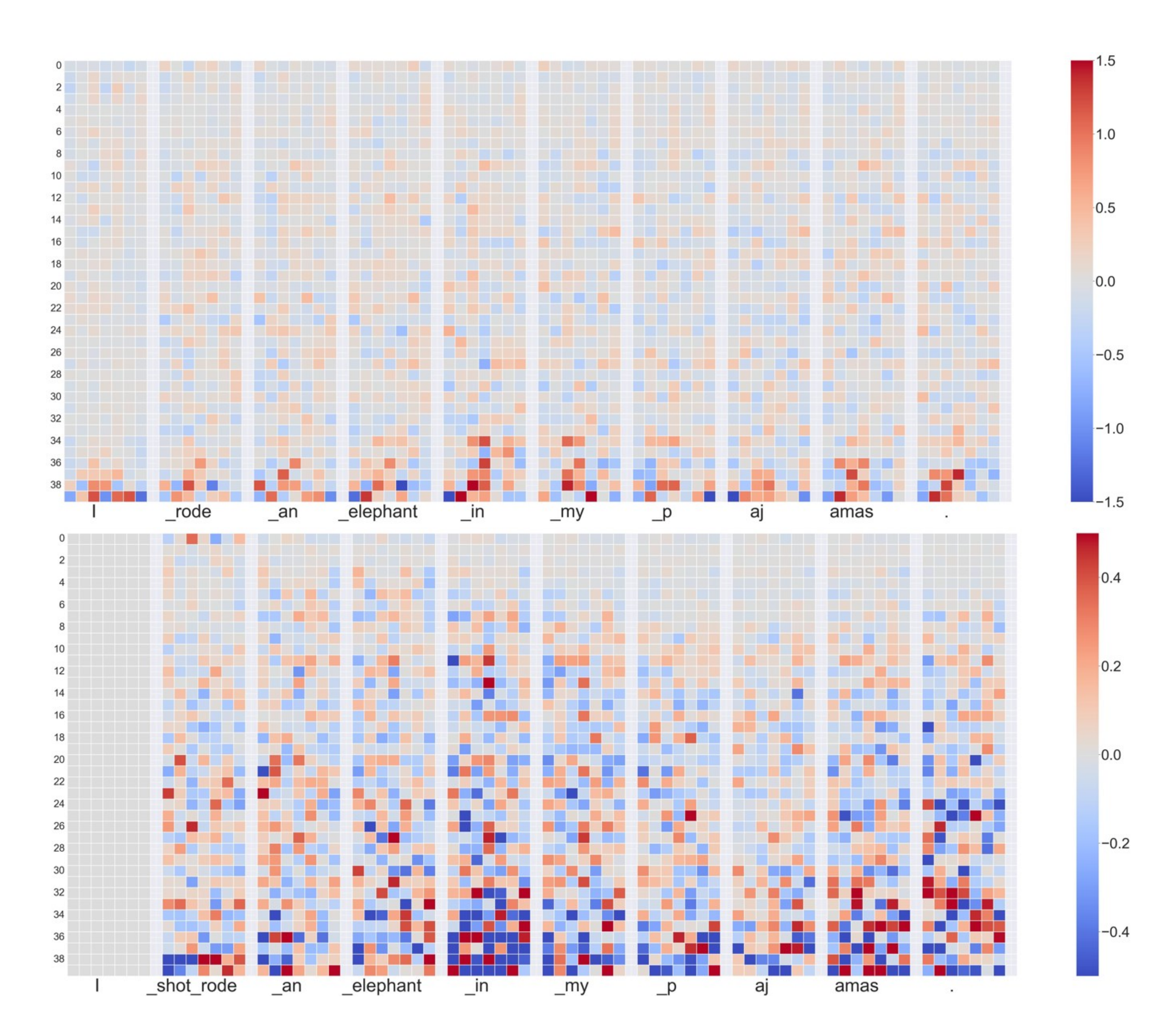



图 14 上图为将例句中的“shot”替换为“rode”后，对应的百川 2 模型内部特征展示；下图为减去$R_0$之后的结果

# 4 总结与展望

大语言模型的快速发展为各类复杂问题提供了新的智能化解决方案，包括帮助人类更好地分析与理解自然语言，这给语言学研究带来了新的思路与方法。本文以自然语言中的歧义现象为例，评估了以 GPT4、百川 2 与 ChatGLM3 模型为代表的大语言模型对于复杂语言现象的感知与分析能力。作为目前最好的通用大语言模型之一，GPT4 可以运用语言学知识，有效融合歧义消解与句法分析两个传统的自然语言处理任务，准确分析给定例句中的歧义现象。相比之下，开源大语言模型百川 2 对于歧义的感知能力较弱，需要通过比较明确的提示引导来实现准确分析。ChatGLM3 的尺寸更小，其对歧义的感知和分析能力则更弱。另一方面，对于开源大语言模型，我们可以方便地获取模型在处理文本时的内部神经元特征。参照神经语言学相关研究方法，我们探究大语言模型内部活动与语言学现象之间的对应关系。由于百川 2 模型对于歧义的感知和分析能力较弱，给定例句在不同的上下文中的特征表示差别较小，但通过不同上下文中模型特征的对比与过滤，我们依然可以明显观察到歧义现象与模型内部特征表示之间的相关性。

本文通过相关实验初步验证了借助大语言模型进行语言学研究的可能性。随着大语言模型，特别是开源大模型本身的进步，其对复杂语言现象的感知能力和分析能力也会越来越强，这也将为我们通过大模型研究各类复杂的语言现象和问题带来更多的可能性，包括利用更大的数据样本，通过更深入地分析实现大语言模型内部特征与语言学理论的对齐，从语言学的角度为大语言模型提供可解释性，促进语言学理论的发展，同时推动大模型相关技术的进步。

# 参考文献

[1] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[3] OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774.

[4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

[5] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., ... & Wu, Z. (2023). Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

[6] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022, May). GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 320-335.

[7] 朱德熙. 句法结构[J]. 中国语文, 1962(8/9): 54-61.

[8] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[9] Phillips, C., & Sakai, K. L. (2005). Language and the Brain. Yearbook of science and technology, 166-169.

[10] Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature reviews neuroscience, 8(5), 393-402.

[11] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., ... & Hendrycks, D. (2023). Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.